

Responsible AI & Risk Plan

1. POSSIBLE AI FAILURE CASES

Five realistic ways Smart Split could fail inside the group-bill-split flow:

- **OCR service outage:** During the 8–10 pm dinner rush the OCR endpoint goes down, so tapping 'Scan Bill' returns nothing and the planner is stuck mid-split with friends waiting.
- **Hallucinated / misread line items:** The model reads Rs 160 Garlic Bread as Rs 760, or invents a line not on the bill, silently inflating one member's share and creating an unfair split.
- **Wrong item-to-person assignment:** NLU maps 'I had the pizza' to the wrong person, or to the wrong line when two pizzas exist, so someone pays for food they didn't eat.
- **Latency spike / context overflow:** A long itemized bill (50+ lines) or peak-hour traffic pushes scan-to-split time past 5 seconds; users abandon and fall back to a calculator.
- **Bill-image privacy leakage:** A photographed bill containing the payer's card last-4 or name is logged, cached, or surfaced inside the shared group thread, exposing financial data.

2. RESPONSIBLE AI RISKS

- **Privacy risk:** Bill photos can carry personal information — name, card last-4, GSTIN. If retained beyond the split or shown in the group thread, the feature exposes data the user never meant to share.
- **Bias & Fairness risk:** OCR trained mostly on English printed bills underperforms on regional-language menus and handwritten bills — exactly the Tier-2 users we serve — producing worse splits for them.
- **Hallucination risk:** A confidently wrong amount makes one friend overpay; because the whole promise is 'fair split,' one visible error breaks the feature's core value.
- **Safety risk:** The harm is financial and social, not physical: a wrong split causes real money loss and friction between friends — the very awkwardness the feature was meant to remove.
- **User-trust risk:** One wrong split early in adoption and users revert to WhatsApp + calculator — the manual behaviour we are replacing — and won't return.
- **Overdependence on AI:** If users stop sanity-checking totals and trust the AI blindly, silent misreads go uncaught and compound across many splits.

3. GUARDRAILS AND MITIGATION PLAN

Risk	Possible Impact	Mitigation / Guardrail
OCR outage	No split; abandonment at peak	Circuit-breaker drops instantly to manual split; auto-retry; status banner.
Hallucinated amount	Overpay; trust break	Hard guardrail: sum of shares MUST equal bill total or send is blocked; per-item confidence; low-confidence items flagged amber.
Wrong assignment	Wrong person charged; conflict	Every item editable; 'review before send' mandatory; unassigned items flagged.
Privacy / image leak	Financial PII exposure	Encrypt in transit and at rest; redact card digits/names pre-processing; raw image never posted to the group; short retention.

Bias on regional bills	Core Tier-2 users underserved	Multilingual OCR; confidence-gated manual fallback; log corrections to retrain.
Over-dependence	Silent errors uncaught	Always show itemised breakdown + AI disclaimer; thumbs up/down to surface bad splits.

4. FALLBACK BEHAVIOR

- **When the AI fails:** The user drops straight into the manual-split UI (equal / by share) with no waiting — File 4, Screen 6. AI scanning auto-retries next time.
- **Graceful degradation:** Low OCR confidence → prompt a retake → if still unreadable, offer manual entry pre-filled with whatever total could be read (File 4, OCR-uncertainty screen).
- **When unavailable / unsure:** A non-blocking banner: 'Smart Split is busy — you can still split manually, no waiting.'
- **If the output is risky (financial):** If shares don't equal the total, the send is blocked with 'these amounts don't add up' — the system never silently sends a wrong split.

5. HUMAN-IN-THE-LOOP PLAN

- **When a human takes over:** A member reports a wrong item twice, the user taps 'Talk to a support agent,' overall confidence stays below 60%, or a receiver disputes their amount.
- **Escalation triggers:** Explicit button/keyword, repeated-correction count ≥ 2 , or the confidence threshold — File 4, Screen 7.
- **Escalation path & speed:** In-app chat routes to a payments support agent who can view the redacted bill image and the AI's extracted items, target first response under 2 minutes; the agent can correct and re-send.
- **Sensitive (financial dispute):** Money disputes route to a payments specialist; suspected fraud or coercion escalates to the risk team, not the AI.

6. AI TRANSPARENCY

- **Labels:** 'Scan Bill (AI)' + 'NEW' badge; the processing screen states 'Powered by PhonePe AI.'
- **Disclaimer:** 'AI-generated split. Review before sending — every amount is editable.'
- **Confidence indicators:** Low-confidence items highlighted amber; if overall confidence $< 60\%$, the user is warned before sending.
- **Feedback mechanism:** 'Was this split correct?' thumbs up/down on the output screen.
- **Escalation path:** 'Talk to a support agent' reachable from the help screen.

7. AI EVALUATION METRICS

Metric	What it measures	Why it matters	Success looks like
Item-extraction accuracy	% of items + prices read correctly	A fair split depends on reading the bill right	$\geq 95\%$ on clear printed bills
Hallucination rate	% of splits with an item/amount not on the bill	Directly causes overpay	$< 1\%$ of splits
Response latency (P95)	Scan-to-split time	Drives mid-flow abandonment	< 5 seconds (P95)

Data-privacy compliance	% images with PII redacted; leak incidents	Regulatory + trust	100% redaction, 0 leaks
Cost per interaction	OCR + compute cost per scan	Unit economics of a free feature	< Rs 0.50 per scan
User satisfaction (CSAT)	Thumbs-up rate / survey	Trust and repeat usage	$\geq$ 85% thumbs-up